# 和歌データベースにおける特徴パターンの発見

竹田正幸[†] 福田智子[††]
南里一郎[†††] 山崎真由美[†]

本研究では，和歌のデータを対象に，歌集の特徴を抽出する問題を扱う．特徴として，「*せば*ざらましを*」などの付属語のパターンを考える．付属語のパターンは，表現技法上の特徴を表しており，たとえば，「*せば*ざらましを*」は反実仮想に対応する．1つの歌集に表れるパターンの数は数十万にものぼるため，そのすべてを研究者が吟味することは，現実的には不可能である．そこで，その大量のパターンの中から「重要」と思われるものだけを，数百程度のオーダーで自動抽出することを考えたい．これが可能となれば，研究者はそれらのパターンを重点的に吟味することにより，有用な知見を得ることができよう．このことを実現するためには，「重要」性を形式的に定義することが必要となる．この定義は非常に難しい問題であるが，本研究では，これを Brāzma ら（1996）にならって，最小記述長（Minimum Description Length; MDL）原理に基づいて与える．この手法を5つの歌集に適用したところ，和歌の研究者にとって有用なパターンが得られることが分かった．また，この経験に基づき，和歌文学研究支援のためのテキストデータマイニングシステムを作成した．このシステムは，研究者を主体とした研究を支えるための有用な道具となりうる．著者らは，このシステムを用いて，新しいスタイルの和歌文学研究を進行中である．

# Discovering Characteristic Patterns from Classical Japanese Poem Database

MASAYUKI TAKEDA,[†] TOMOKO FUKUDA,[††] ICHIRO NANRI[†††] and MAYUMI YAMASAKI[†]

WAKA is a form of traditional Japanese poetry with a 1300-year history. In this paper, we attempt to discover characteristics common to a collection of WAKA poems. As a formalism for characteristics, we use regular patterns where the constant parts are limited to sequences of *adjuncts*, i.e., auxiliary verbs and postpositional particles. For example, we consider patterns such as *SEBA*ZARAMASHIO*, where SEBA is a chain of an auxiliary verb SE and a postpositional particle BA, and ZARAMASHIO is a chain of two auxiliary verbs ZARA and MASHI and a postpositional particle O. This pattern corresponds to the subjunctive mood. We call such patterns FUSHI. The problem is to find more or less automatically significant FUSHI patterns that characterize the poems. Solving this problem requires a reliable significance measure for the patterns. Brāzma et al. (1996) proposed such a measure according to the Minimum Description Length (MDL) principle, in which the most significant pattern set is the one that minimizes the sum of the length of the patterns and the length of the strings when encoded with the help of the patterns. Using this method, we report successful results in finding patterns from five anthologies. Some of the results are quite stimulating, and we hope that they will lead to new discoveries in Japanese literature. Based on our experience, we also propose a pattern-based text data mining system that consists of a pattern matching part and a pattern discovery part. Further research into WAKA poetry is now proceeding using this system.

## 1. はじめに

近年，人文科学系の様々な分野において計算機を用いた研究が行われるようになってきた．国文学の分野も例外ではない．計算機を用いた研究の前提として，研究資料のデータベース化が不可欠であるが，国文学研究においては，本文テキスト，研究論文などの1次情報や，抄録，目録，索引などの2次情報に加えて，写

† 九州大学大学院システム情報科学研究科
Department of Informatics, Kyushu University
†† 福岡女学院大学
Fukuoka Jo Gakuin College
††† 純真女子短期大学
Junshin Women's Junior College

本，版本などの0次情報が必要であり[1]，そのデータベース化の作業は，地道で多大な労力を要する．現在，国文学研究資料館を中心としてデータベース化の作業とそのために必要な研究開発が進められている[2)~4)]．また，研究者個人レベルで入力したデータを共有化しようという動きもさかんであり，情報処理語学文学研究会（JALLC）や勉誠データセンターなどに集積され，草の根的にデータベース化が進みつつある．

ところで，国文学研究者の多くは，これまで研究における計算機利用に関してかなり否定的であった．これは，肯定するにせよ，否定するにせよ，計算機の役割を「打ち出の小槌」のようなイメージでとらえていたためであろう．しかし，ワープロの普及などを1つの契機として，単なる道具としての計算機のイメージが浸透しつつある．このことを端的に表しているのが，「コンピュータに文学が分かるものか！」といった否定的反応に対し，「それでは，あなたの万年筆に文学が分かりますか?」と切り返すという有名な話である．研究の主体は，あくまで人間であって，計算機ソフトウェアはそのための有用な道具でなければならない．

現在のところ，国文学研究における計算機の利用法は，万年筆の延長としてのワープロソフトの利用を別にすると，主として次の2つである☆．

- エディタの機能やデータベースソフトを用いた索引作成．
- 全文データを対象としたキーワード検索による用例収集．

これらの利用法は非常に有用であり，従来人手で行っていた単純作業から研究者を解放し，より創造的な仕事へ専念させるものである．しかし，国文学研究者の多くは，これ以上の利用法の可能性に関して，必ずしも肯定的ではない．計算機は「たかが道具にすぎない」が，その機能は固定的なものではなく，ソフトウェアによって変わるものである．それでは，どのような道具であれば，国文学研究にとって有用であるのだろうか．この問題の解を得るためには，国文学研究の現場と密接な関係を保ちながら，あくまで研究者を主体とした研究のための，より高度な道具としての在り方を探っていく必要がある．

本研究では，和歌のデータを対象に，歌集の特徴を抽出する問題を扱う．特徴として，「*せば*ざらましを*」などの付属語のパターンを考える．付属語のパターンは，表現技法上の特徴に対応するものであり，たとえば，パターン「*せば*ざらましを*」は反実仮想という表現技法に対応している．

和歌のデータが機械可読化されていれば，パターンに合致する和歌をすべて取り出すことは容易である．あるパターンに着目してその用例を調査し，意味のある結果が得られれば，それは研究成果となる．これまで，そのようなパターンは，研究者が任意に与えるしかなかった．だが，もしそれを計算機を利用することにより自動的に抽出できれば，新たな発見への端緒となることも期待できるのである．

5章で示すように，1つの歌集に表れるパターンの異なり数は数十万にものぼるため，そのすべてを研究者が吟味することは，現実には不可能である．そこで，その大量のパターンの中から「重要」と思われるものだけを，数百程度のオーダーで自動抽出することを考えたい．これが可能となれば，研究者はそれらのパターンを重点的に吟味することにより，有用な知見を得ることができよう．

このことを実現するためには，「重要」性を形式的に定義することが必要となる．この定義は非常に難しい問題であるが，本研究では，これをBrāzmaら[6)]にならって，最小記述長（Minimum Description Length; MDL）原理[7)]に基づいて与える．理想的には，この基準に従って抽出したパターンすべてが「重要」であり，かつ「重要」なパターンすべてが抽出されることが求められるが，もちろんそれは現実的ではない．その意味では，データマイニングのようなアプローチの仕方が鍵となるであろう．データマイニングとは，大量のデータの集積から，その中に潜む自明でない規則を半自動的に発掘する技術をいう．データマイニングは，実用的見地から注目を集めビジネスや科学技術分野における応用を目指してさかんに研究されている[8)]が，その成功の理由は，有用な情報だけを発掘することを目指さずに，発掘した情報を評価してその中から有用な情報を取り出す仕事をユーザに委ねた点にある．国文学研究においても，発掘した結果を文学的立場から意味付けし評価していく仕事は，研究者に委ねるべきであるので，このようなアプローチの仕方には期待が持てる．そこで，著者らは，和歌文学研究支援のためのテキストデータマイニングシステムの開発を目指す．このようなシステムは，上に述べた，研究者を主体とした研究を支える，より高度な道具となりうるものである．近年，角川書店より，約45万首もの和歌を収録した新編国歌大観のCD-ROM版が発売され，また，国文学研究資料館が独自に翻刻から行った二十一代集の検索サービスがWWW上で提供されるなど，和歌データの機械可読化が進んでいることを考えると，

☆ このほか，統計的手法に基づく計量的研究[5)]も行われている．

このようなツールの開発によって，和歌文学研究が飛躍的な発展を遂げることが期待できよう．

本研究は，情報科学の研究者（第1，4著者），和歌文学の研究者（第2著者）および国語学の研究者（第3著者）が共同して行った．なお，本論文の一部は，文献9)に発表した内容に基づいている．

## 2. 和歌の特徴

### 2.1 歌語による特徴づけ

和歌の特徴といえば，これまでもっぱら歌語に着目した研究が行われてきた．ここで，歌語とは，和歌に用いられた名詞，動詞，形容詞などの自立語を指すものとする．たとえば，以下に示す3首に共通している歌語は，「しらつゆ」「たま」「いと」であるということになる．

あさみどり いと よりかけて しらつゆ を
たま にもぬける 春の柳か
（古今集27番）

秋ののに おく しらつゆ は 玉 なれや
つらぬきかくる くもの いと すぢ
（古今集225番）

白露 を 玉 にぬくとや ささがにの
花にも葉にも いと をみなへし
（古今集437番）

そして，「たま」のような「しらつゆ」を「いと」で貫くという，同様の見かたによって，各々の歌が仕立てられる．このように，共通した歌語を含む歌は，おおむね表現世界も類似していると考えることができる．

歌語に着目した特徴づけは，和歌が機械可読化され，計算機によって処理可能となるかなり以前から，人手で作成した歌語索引などにより，さかんに行われてきた．そもそも，和歌の表現世界のイメージは，はっきりとした形で研究者の記憶に残りやすいため，内省によって，表現世界の類似した歌に気づくことができる．すなわち，研究者の〈読む〉という行為になじみやすい特徴づけであるといえる．

しかし，歌語のみによる特徴づけには，限界がある．たとえば，紀貫之が桜の花を多く詠んだからといって，「貫之は桜の花を好んだ」と結論するのは，短絡に過ぎるであろう．というのは，当時は詠歌に際して題が与えられることがあり，歌人が歌語を任意に選択できない場合があったと考えられるからである．

### 2.2 ふしによる特徴づけ

それでは，和歌の特徴として，他にどのようなものを考えればよいだろうか．古今集169番の歌を例にとって見てみよう．

あききぬと めにはさやかに 見えねども
風のおとにぞ おどろかれぬる
（古今集169番）

これは，秋の到来を風の音で知るという，秋部冒頭の有名な歌である．この歌と，次の2首を比べてみると，歌語の支配する表現世界がまったく異なるにもかかわらず，ある類似性が存在することに気づく．

せきとめて うぢのかはなみ よせねども
つきにそひてぞ 心ゆきぬる
（為仲集97番）

松しまの あまのとまやは しらねども
我が袖のみぞ しをれわびぬる
（後鳥羽院御集1041番）

すなわち，「ねども」「ぞ」「ぬる」という付属語もしくは付属語の列が，同じ順序で用いられているため，これらの歌が似ていると認識されるのである．本論文では，和歌の骨格をなすこのような構造をふし（節）とよび，

*ねども*ぞ*ぬる*

のように表すことにする．

和歌は，歌語という「素材」を，ふしという「器」に盛りつけたものととらえることができる．つまり，ふしは和歌の表現技法に関する特徴と考えられる．素材（歌語）の選択が，ある程度制限されていたとすれば，歌人たちの関心は，それをどのような器（ふし）に盛るかという点に集中したに違いない．

先に述べたとおり，歌語は，基本的に自立語である．一方，ふしは，付属語のなすパターンである．付属語に着目した和歌の特徴づけの研究は，これまでほとんど行われていない☆．その理由の1つとして，研究者が和歌を〈読む〉際に，付属語はそれ自身実質的な意味を持たないため，イメージを喚起せず，記憶に残りにくいことがあげられよう．つまり，内省によって付属語に関する特徴を抽出することは，かなり困難なのである．そこで，研究者は，〈読む〉という姿勢をいったん脇に置き，膨大な数の和歌に，片端から目を通す作業の必要に迫られる．特定のふしを含む歌の検索は，計算機を用いれば容易に行える．ところが，検索すべきふしは，付属語のなすパターンであるから，その数は組合せ的に膨大であり，考えうるすべてのふしについてこの作業を行うのは，ほぼ不可能に近い．

そこで本研究では，和歌の集合に表れるパターンのうち，「重要」と思われるものだけを計算機を用いて

☆ 国語学の視点からは，付属語に関する語彙索引の作成なども行われているが，国文学の視点からの研究はほとんど行われていない．

抽出し，それらを重点的に吟味することにより，表現技法に関する特徴を獲得することを目指す．このような研究は，計算機を用いてはじめて可能となるといえよう．

### 2.3 ふしと言語観

ふしを，表現技法上の特徴であると主張するからには，パターンの定数部分に表れる文字列が何であるかを明らかにし，それから作られるパターンが全体としてどう機能するかについての見通しを持っておかねばならない．これを示すことにより，その背景にある言語観も，おのずと明らかになることであろう．

さて，前節までは，ふしを「付属語」のなすパターンとして述べてきたが，学校文法に則った「自立語」「付属語」という単純な枠組みでは，どうやら和歌の実際をとらえることはできないようである．これについて，具体例に即して検討する．

まず，次の3つの歌をみてみよう．

月見 れば ちぢに物 こそ かなし けれ
わが身ひとつの 秋にはあらねど
（古今集 193 番）

ゆふさ れば わが身のみ こそ かなし けれ
いづれの方に 枕さだめむ
（後撰集 739 番）

老いぬ れば おなじこと こそ せられ けれ
きみはちよませ きみはちよませ
（拾遺集 271 番）

これらの3首には，

*れば*こそ*けれ*

というパターンが共通して表れている．一見して，前節で示したふしの例と大差はないように思われる．しかしここで，パターンの定数部分に注目してみると，「ば」「こそ」はいずれの歌においても助詞であるのに対し，「れ」「けれ」は，歌によって文法的範疇が違っていることに気づく．すなわち，1，2首目にある「れば」の「れ」は動詞の活用語尾であるが，3首目のそれは助動詞の語尾である．また，1，2首目の「けれ」は形容詞の活用語尾であるが，3首目のそれは助動詞である．このように，同一のパターンのように見えても，その定数部分の文法的範疇は異なっている．

それでは，この3首において，同一のパターン「*れば*こそ*けれ*」が生起しているとはいえないのであろうか．著者らは，そうではなく，これも同一のパターンであると考える．というのは，当時の歌人たちの意識や言語感覚では，これら3首が同一のパターンを含んだものに見えていたと考えられるからである．ここでは，文法的範疇はさておいて，文字列としてはまったく同じである，ということを重視すべきであろう．

もちろん，自立語の活用語尾は，あくまでも自立語の一部分であって付属語ではない．しかし，膠着語といわれる日本語の特徴を考えるとき，活用語尾には，後続の語との関係を規定するという点で，付属語と共通する文法的性質がある．したがって，動詞や形容詞の活用語尾も，付属語的なものとして，ふしを構成する要素と考えるのが妥当であろう．そもそも助動詞には動詞を起源とするものも多く，文字列として見た場合，ある動詞の活用語尾とまったく同一であることが少なくない．形容動詞の活用語尾に至っては，指定の助動詞「なり」と組成が同じであるため，文字列としてつねに一致する．このような事実は，著者らの考え方を基底から支えるものである．

また，自立語の語尾に限らず，自立語それ自体がふしを構成する要素となる場合がある．

かつ越えて わかれも行くか あふさかは
人だのめなる 名にこそありけれ
（古今集 390 番）

この和歌では，「越え」「わかれ」「あふさか」「人だのめ」「名」などの自立語が，いわゆる歌語として，一首の表現世界を規定している．では，ふしと考えられるのはどこか．そこで我々の目には，「*は*にこそありけれ*」というパターンが浮かび上がってくる．このパターンにある文字列「にこそありけれ」に含まれる「あり」は，学校文法では補助動詞と説明される．動詞であるからには自立語に違いないが，存在を表す「有り」の実質的意味を失い，付属語的要素となっている．この「あり」は付属語の列の中に組み込まれて，「にこそありけれ」全体で1つの働きをなしていると考えられる．つまり，歌語は自立語であるといえそうだが，自立語ならば歌語であるとは必ずしもいえない．自立語であっても，表現世界の形成に寄与しないという意味で，歌語にはなりえず，むしろ，ふしの構成要素となる場合が少なからずある，と考えられる．

以上のように，著者らは，学校文法の概念から離れて，ふしというものを考える．ふしの定数部分となる文字列の文法的範疇は，実に雑多なものになるであろう．もちろん付属語（助詞・助動詞）が中心となるが，活用する自立語の語尾をも含み，場合によっては，名詞や動詞，副詞などの自立語自身さえも含むことがあると思われる．しかし，そうした文法的範疇の違いも，ひとまとまりの文字列として見ることにより，すべて捨象される．このように考えた文字列を，ここで著者らは，あらためて**付属語列**と呼ぶことにする．

付属語列の果たす機能は，それを構成する各々の要素の機能の単なる総和として得られるものではない．したがって，付属語列を分割不能な文字列として扱い，個々の付属語列の果たす機能を考察していく必要がある．そして，付属語列のなすパターンであるふしは，パターン全体として1つの機能を果たしている．その機能は，和歌の構造や韻律と密接に関係していると考えられる．

### 2.4 和歌とふしの関係

1つの和歌に対して，そのふしは一意に定まるものであろうか．もし，そうであれば，本研究の目的である歌集からの特徴抽出は，単に，歌ごとのふしを集めて，その頻度を調べる程度のことになる．しかし，与えられた歌に対して，ふしは一意に定まらない．このことを，以下の例でみてみよう．

古今集420番に，菅原道真の有名な歌がある．

このたびは ぬさもとりあへず たむけ山
紅葉の錦 神のまにまに
（古今集420番）

後世，これをもとにした歌がいくらか詠まれることになるのだが，その踏まえ方は様々である．たとえば，「*あへず*のまにまに*」というパターンを取り込んだ歌として，以下に示す3首があげられよう☆．

みるままに 筆もとりあへず 手向山
このことのはも 神のまにまに
（柏玉集1901番）

かつこえて ぬさもとりあへず あふ坂の
花のしらゆふ せきのまにまに
（道助法親王家五十首216番）

とりあへず ぬさと手向けし 神にまた
ちるもみぢばを 風のまにまに
（黄葉集913番）

ところが，同じ道真の歌に対し，「*この*は*あへず*」という箇所に視点をずらすと，上に列挙した歌とはまったく別に，次のような歌が浮かび上がってくる[10]．

このたびは えだにぬひあへず 唐衣
たつにとまらぬ なみだならぬに
（元真集190番）

このように，同じ歌であっても，視点の違いや比較する歌によって，ふしとすべき箇所は異なってくるのである．この事実は，5.3節で述べるように，パターン抽出アルゴリズムの評価を行うための「正解」となるデータを作成できないことを意味している．

☆「あへず」は，動詞「あふ」と助動詞「ず」の連接であり，「あへず」全体で「~しきれない」の意で機能している．一方，「のまにまに」については，諸説あるが，助詞「の」，名詞「まにま」，助詞「に」の連接であろう．「まにまに」は連体修飾語を承けて連用修飾句を構成する機能を持つ．

## 3. パターン抽出へのアプローチ

歌集からふしとよぶパターンを抽出するためには，以下の2つが必要である．

- 和歌において，付属語列を決定すること．
- 歌集からふしとよぶパターンを抽出すること．

この章では，この2つの問題に対するアプローチの方法について論じる．

### 3.1 付属語列の決定

付属語列の決定のためには，和歌に形態素解析に類した処理を施せばよいが，その過程で曖昧さの問題が生じる．現代日本語文に対する形態素解析の研究は古くから行われているが，決定的な方法はいまだに得られていない．まして対象は古文である．古文の形態素解析については少数の研究がある[11]が，辞書などの未整備もあって実用レベルにはほど遠い．そこで，本研究では，通常の形態素解析を断念し，付属語列と同じ形の文字列は，付属語列と見なすことにする．

こうすることには利点がある．というのは，形態素解析を用いた場合には，パターン抽出のステップで生じた問題点が，形態素解析の解析ミスに起因している可能性も考えなければならないが，一般に，解析ミスの種類やその原因は単純ではないため，問題の所在が不明確になるからである．

### 3.2 パターンの抽出

文字列の集合からパターンを抽出する研究は，機械学習の分野において古くから行われている．パターン言語の帰納推論[12]においては，言語同定のモデルとして極限における同定（Identification in the limit）[13]が用いられる．このモデルでは，あらかじめ設定した言語クラスの中に学習すべき言語が属していることを前提とし，その言語に属する語（および属さない語）を，正例（および負例）として無限に与え続ける機構が仮定される．この学習の過程において最終的に仮説が正しい解に収束することが要求されるが，言語をいつ同定したかを判定する必要はない．また，学習の過程で中間的に出力される仮説の精度は問われない．一方，PAC学習[14]においては，学習に必要な例の数の上限が議論される．しかし，我々の場合には，学習に十分な数の例が与えられる保証はない．与えられるのは，歌集という限られたサイズの正例の集合のみである．

本研究においては，抽出すべきパターンは，それぞれ，表現技法に対応している．すなわち，ある時代に用いられた表現技法，あるいは，ある歌人が用いた表

現技法としてのパターンを抽出することが，本研究の目的である．したがって，学習すべき言語のクラスとしては，パターン言語の有限和のクラス[15]を考える必要がある．正例の集合としては，特定の時代に編まれた歌集，あるいは，特定の歌人の私家集を用いることが考えられよう．しかし，負例に関しては，それを得るための有効な手段がない．したがって，有限個の正例のみからパターン言語の有限和を学習する問題を考えることになる．この場合，過剰な一般化や，過剰な具体化が問題となる．すなわち，極端にいえば，正例のすべてを1つのパターン「$*$」で覆うことも可能であるし，逆に，正例の各々に対して1つのパターンを用いることも可能である．だが，実際に必要なのは，その中間的なパターンの集合である．Arimuraら[16]は，正例の集合 $S$ に対する $k$ 極小多重汎化（$k$-minimal multiple generalization; $k$-mmg）を，$S$ を覆うたかだか $k$ 個のパターンから成る極小な集合と定義し，この $k$ 極小多重汎化が，極限における同定に基づいた正例からの帰納推論の観点から最適であることを示した．しかし，この方法を適用しようとすると，次のような問題に直面する．(a) 前もって $k$ の値を見積もる必要があること．(b) $k$ 極小多重汎化を求める多項式時間アルゴリズムは，$k$ を少し大きくとると計算時間が現実的でなくなること．(c) 正例の集合に対し，$k$ 極小多重汎化は一意に定まらないが，どれを選ぶべきかの基準がないこと．

一方，Brāzmaら[6]は，別の観点から最適な被覆の定義を与えた．すなわち，正例の集合 $S$ の最適な被覆とは，ある単純な確率モデルのもとで，条件付き確率 $Pr(\Pi|S)$ の値を最大にするパターンの集合 $\Pi$ として定義される．この基準は，最小記述長原理[7]によるものと等価である．この手法は，DNA の塩基配列の集合から PROSITE パターンを抽出する問題に適用され，その有効性が示されている[17]．

一般に，望ましいパターンとして以下の基準が考えられる．

(1) できるだけ多くの正例に共通すること．
(2) できるだけ多くの定数記号を含むこと．
(3) できるだけ $*$ の数が少ないこと．

Brāzmaらの基準は，これら相反する3つの間で適当にバランスをとったものであるといえる．そこで，本論文ではこの手法を適用することにした．

## 4. MDL 原理に基づく最適な被覆

この章では，Brāzmaら[6]に沿って，MDL 原理に基づく最適な被覆の定義を与え，最適解を近似するアルゴリズムを示す．

### 4.1 被覆の最適性の定義

$\Sigma$ を文字の有限集合とし，$* \notin \Sigma$ をギャップ記号（gap symbol）とする．**パターン**とは，$\Sigma \cup \{*\}$ を字母とする長さ1以上の記号列をいう．文字列 $w \in \Sigma^+$ とパターン $\pi \in (\Sigma \cup \{*\})^+$ に対し，$\pi$ 中の $*$ の出現を，それぞれ，空でない文字列で置き換えて $w$ が得られるとき，**$w$ は $\pi$ に合致する**という．パターン $\pi$ に合致する文字列全体の集合を，**$\pi$ によって定義される言語**といい，$L(\pi)$ で表す．

さて，パターン

$$\pi = *\beta_1 * \cdots * \beta_h * \quad (\beta_1, \ldots, \beta_h \in \Sigma^+)$$

と $B \subseteq L(\pi)$ となる文字列の集合

$$B = \{\alpha_1, \ldots, \alpha_n\}$$

を考えよう．このとき，集合 $B$ は，パターン $\pi$ と以下の文字列群によって記述できる．

$$\begin{array}{cccc} \gamma_{1,0} & \gamma_{1,1} & \cdots & \gamma_{1,h} \\ \gamma_{2,0} & \gamma_{2,1} & \cdots & \gamma_{2,h} \\ & & \vdots & \\ \gamma_{n,0} & \gamma_{n,1} & \cdots & \gamma_{n,h} \end{array}$$

ここで，各 $i = 1, \ldots, n$ について，

$$\alpha_i = \gamma_{i,0}\beta_1\gamma_{i,1}\cdots\gamma_{i,h-1}\beta_h\gamma_{i,h}$$

である．集合 $B$ のこのような記述法を，**パターン $\pi$ を用いた記述法**とよぶことにする．

パターンや $*$ に代入する文字列の生起に関して無記憶情報源モデルを仮定し，これらを，ハフマン符号のような文字単位の符号を用いて記述することを考えよう．その際の文字列 $\alpha$ の記述長を $\|\alpha\|$ で表すことにする．簡単のため，文字列間の区切り文字などを無視して考えると，パターン $\pi$ を用いた記述法における集合 $B$ の記述長は，次のようになる．

$$\|\pi\| + \sum_{i=1}^{n}\sum_{j=0}^{h}\|\gamma_{i,j}\|$$

パターン $\pi$ 中の $*$ をすべて取り除いた文字列を $c(\pi)$ で表すと，次を得る．

$$\|\alpha_i\| = \sum_{j=0}^{h}\|\gamma_{i,j}\| + \|c(\pi)\|$$

そこで，集合 $B$ の記述長は，

$$\|\pi\| + \sum_{i=1}^{n}(\|\alpha_i\| - \|c(\pi)\|)$$

$$= \sum_{i=1}^{n}\|\alpha_i\| - \Big(\|c(\pi)\| \cdot |B| - \|\pi\|\Big)$$

となる．ここで，$|B|$ は集合 $B$ の大きさを表すものとする．

$A$ を文字列の有限集合とする．パターンと $A$ の部分集合の対の有限集合

$$\Omega = \{(\pi_1, B_1), \ldots, (\pi_k, B_k)\}$$

で以下を満たすものを $A$ の被覆とよぶ．

- $B_j \subseteq L(\pi_j) \quad (j = 1, \ldots, k)$.
- $A = B_1 \cup \cdots \cup B_k$.
- $B_1, \ldots, B_k$ は互いに素．

各 $j = 1, \ldots, k$ について，集合 $B_j$ を $\pi_j$ を用いて記述するとき，集合 $A$ の記述長は，

$$M(\Omega) = \sum_{\alpha \in A} \|\alpha\| - C(\Omega)$$

となる．ここで，$C(\Omega)$ は集合 $A$ の記述長に関する利得であって，以下で与えられる．

$$C(\Omega) = \sum_{j=1}^{k} \Big( \|c(\pi_j)\| \cdot |B_j| - \|\pi_j\| \Big)$$

さて，$M(\Omega)$ を最小にする $\Omega$ またはそのパターンの集合を，集合 $A$ に対する最適な被覆と定義する．$\sum_{\alpha \in A} \|\alpha\|$ は集合 $A$ にのみ依存するので，$M(\Omega)$ を最小化する問題は，$C(\Omega)$ を最大化する問題と等価である．

### 4.2 符号化の詳細

上で与えた最適な被覆の定義は，パターンや * に代入する文字列の符号化の方法に依存する．そこで，どのような符号化を選ぶかという問題が新たに生じる．文献 6) の符号化では，パターンおよび * に代入する文字列は，区切り記号とともに，ある確率分布のもとでの文字単位の最適符号によって符号化される．したがって，$C(\Omega)$ の式は，区切り記号や * の生起確率をパラメータとして含む．

しかし，和歌の場合は，ほとんどが 31 文字であるため，符号化すべき文字列は長さが $m = 32$ 未満と考えてよい．そこで，以下に述べるような素朴な方法を選んだ．文字列 $w \in \Sigma^*$ を，$w$ の長さと $\Sigma$ 上の確率分布 $P$ のもとでの最適符号によって $w$ の各文字を符号化したビット列との対で表す．実用的には，$P(a)$ の値をデータベースにおける文字 $a \in \Sigma$ の相対頻度とする．文字列 $w$ を符号化したビット列の長さを $\ell_P(w)$ で表す．パターン $\pi$ における * の生起回数を，$n_*(\pi)$ で表す．また，$|w| < m$ なる正数 $m$ を仮定する．すると，$C(\Omega)$ は次のようになる．

$$C(\Omega) = \sum_{j=1}^{k} \Big( u(\pi_j) \cdot |B_j| - v(\pi_j) \Big)$$

ここで，

$$u(\pi) = \ell_P(c(\pi)) - n_*(\pi) \log_2 m + \log_2 m$$
$$v(\pi) = \ell_P(c(\pi)) + n_*(\pi) \log_2 m$$

である．

### 4.3 Brāzma らの近似アルゴリズム

与えられた文字列の集合に対して最適な被覆を求める問題は，最小集合被覆問題 [18] をその特別な場合として含むため，NP 困難である．そこで，Brāzma ら [6] は，次のように問題を設定し，この問題に近似解を与える貪欲（greedy）アルゴリズムを示した．

> 文字列の有限集合 $A$ とパターンの有限集合 $\Delta$ が与えられたとき，$\Delta$ の元をパターンとするような $A$ の被覆 $\Omega$ で，$M(\Omega)$ を最小にする $\Omega$ を求めよ．

Brāzma らの近似アルゴリズムを**図 1** に示す．このアルゴリズムは，while ループの 1 回の繰返しごとに，

$$u(\pi) - \frac{v(\pi)}{|L(\pi) \cap U|}$$

の値が最大となる $\pi$ を選んで被覆の集合 $\Omega$ へ加えるものである．ここで，$U$ は各時点においてそれまでに選んだパターンのいずれによっても覆われていない $A$ の元の集合である．このアルゴリズムで得られる近似解に対する $M(\Omega)$ の値は最適解の場合のたかだか $\log_2 |A|$ 倍であることが保証される．また，アルゴリズムの計算時間については，集合 $\{(\pi, L(\pi) \cap A) \mid \pi \in \Delta\}$ の計算に $O(\sum_{\pi \in \Delta} |\pi| + |\Delta| \cdot \sum_{\alpha \in A} |\alpha|)$ 時間を要し，それ以外の部分に，$O(|\Delta| \cdot |A| \cdot \log_2 |A|)$ 時間を要する．

## 5. 歌集からのパターン発見

この章では，歌集からふしを抽出する実験について述べる．上で述べた近似アルゴリズムを適用するためには，付属語列の誤認を減らすことと，付属語列の集合を与えることが必要であるが，これらについては，5.1 節と 5.2 節で述べる．また，5.3 節では，実験の結果について述べる．

### 5.1 付属語の誤認の低減

3.1 節で述べたように，本論文では，付属語列と合致する文字列は付属語列と扱うため，誤認の問題が生じる．この問題を軽減するため，次のような方法を用いた．

- 和歌の句末に生起する付属語列だけに制限した．もちろん，句の途中にも付属語列は生起するが，ふしに関与する重要な付属語列の多くは句末に生起する．
- 和歌には漢字と平仮名が混在しているが，付属語列は基本的に平仮名で書かれている．データには，

**Input:** A finite set $A$ of strings and a finite set $\Delta$ of patterns with $A \subseteq \bigcup_{\pi \in \Delta} L(\pi)$.
**Output:** An approximate solution $\Omega$ in which patterns are chosen from $\Delta$.
**Method:**
**begin**
  $U \leftarrow A$;
  $\Omega \leftarrow \emptyset$;
  $\Gamma \leftarrow \{(\pi, L(\pi) \cap A) \mid \pi \in \Delta\}$;
  **while** $U \neq \emptyset$ **do begin**
    find $(\pi, F) \in \Gamma$ maximizing $u(\pi) - \dfrac{v(\pi)}{|F \cap U|}$;
    $\Omega \leftarrow \Omega \cup \{(\pi, F \cap U)\}$;
    $U \leftarrow U - F$;
    $\Gamma \leftarrow \Gamma - \{(\pi, F)\}$
  **end**
**end.**

図 1 Brāzma らの近似アルゴリズム
Fig. 1 Approximation algorithm by Brāzma, et al.

すべての和歌について，清音表記された読み仮名が付与されている．いま，ある歌の句の読み仮名が，別の歌の句の読み仮名と等しいとしよう．このとき，2 つの句を等価であると考え，句末の平仮名列の短い方をとることにすれば，付属語列の誤認はいくらか低減されるであろう．そこで，読み仮名の等しい句のうち，句末の平仮名列が最短の句を標準形とよぶことにし，すべての和歌の句を標準形に置き換えたデータを用いることにした．以上の方法を用いても，付属語列の誤認のすべてを除去することはできないが，重要なパターンだけを抽出する目的からは，十分であると考えられる．

### 5.2 付属語列の収集

本論文の設定では，ふしとよぶパターンのクラスは，付属語列の集合を与えることによって定義される．すなわち，$C \subset \Sigma^+$ を付属語列の集合とするとき，ふしとは，以下の形に制限されたパターンである．

$$*\beta_1 * \cdots * \beta_h * \quad (h > 1, \beta_1, \ldots, \beta_h \in C)$$

したがって，ふしを扱うためには，その定数部分となる付属語列の集合を定めておく必要がある．2.3 節で述べたように，本論文で用いる付属語列という術語は，学校文法の範囲を逸脱しており，「のまにまに」のように自立語を含むものもある．このような自立語をすべてあげることは難しく，したがって，付属語列を網羅的に収集することは，容易な作業ではない．そこで，とりあえず，学校文法に沿った範囲で付属語列の集合を定め，これを用いて実験を行うことにした．ただし，活用語の活用語尾は付属語と同様に扱う．以下に，ここで与えた付属語列の形式定義の概略を述べる．

付属語列は，次の 3 つがこの順に並んだものである．

- 活用語（動詞・形容詞・形容動詞）の活用語尾．
- 助動詞の列．
- 助詞の列．

ただし，それぞれ空語の場合を許す．活用語尾，助動詞，助詞の連接には，統語的制約と意味的制約がある．統語的制約は，比較的単純であって，語自身とその直前の語の活用形との組合せに依存する．このような制約を記述するのは容易である．一方，意味的制約を完全に記述するのは，きわめて困難である．ここでは，助動詞間の連接と助詞間の連接についてのみ，意味的制約を与えた．このために，岩波古語辞典巻末の「基本助動詞解説」に沿って，助動詞を 5 つのカテゴリに分類し，その分類に基づく規則を与えた．また，助詞に関しても 6 つのカテゴリに分類し，同様に規則を与えた．

以上のようにして，実験で用いる付属語列の集合 $C$ を定めた．この形式的定義は，「どんぶり勘定」的なものであって余計なものも含むが，ここでは漏れのないことを重視した．また，ここで定義された $C$ は無限集合であるが，このうち，新編国歌大観の和歌のうち一部を除いた 351,390 首の句末に生起した付属語列の異なり数は，29,757 であった．

### 5.3 5 つの歌集に対する実験結果とその評価

パターン抽出法の有効性を検証するためには，歌集の各々の和歌について，ふしをタグ付けしたコーパス（tagged corpus）を人手により作成しておき，これを用いることによって，パターン抽出法の精度を評価する方法が考えられる．ところが，2.4 節で述べたように，和歌に対してふしを一意に決定することができないため，このようなコーパスを作成することができな

表 1 5 つの歌集
Table 1 Five anthologies.

| 歌集名 | 説明 | 歌の数 |
|---|---|---|
| 古今集 | 勅撰集（ 922 年成立） | 1,111 |
| 新古今集 | 勅撰集（1205 年成立） | 2,005 |
| 壬二集 | 私家集：藤原家隆（1158–1237） | 3,201 |
| 拾遺愚草 | 私家集：藤原定家（1162–1241） | 2,985 |
| 山家集 | 私家集：西行法師（1118–1190） | 1,552 |

表 2 5 つの歌集に対する被覆
Table 2 Coverings of five anthologies.

| 歌集 | 生起パターン数 | | 被覆のサイズ |
|---|---|---|---|
| 古今集 | 164,978 | (8,265) | 191 |
| 新古今集 | 233,187 | (12,449) | 270 |
| 壬二集 | 187,014 | (16,425) | 369 |
| 拾遺愚草 | 214,940 | (14,365) | 335 |
| 山家集 | 279,904 | (12,963) | 232 |

い．そこで，以下の点について評価を行うことにした．

(1) 得られたパターンの数は十分少ないか．
(2) 典型的と思われるパターンが抽出できているか．
(3) 自明でないものが抽出できているか．
(4) 歌集ごとの差異が出ているか．

実験には，古今集，新古今集，壬二集，拾遺愚草，山家集の 5 つの歌集を用いた（**表 1**）．最初の 2 つは勅撰集，すなわち，天皇の命令によって編纂された歌集である．古今集は 922 年，新古今集は 1205 年に成立した．そこで，もし，これらの歌集に何か差があるとするならば，それは時代の違いによるものではないかと考えられる．一方，残りの 3 つは，ほぼ同時代に活躍した歌人である藤原家隆（1158–1237），藤原定家（1162–1241），西行法師（1118–1190）の 3 人による私家集である．そこで，3 つの歌集の違いは，歌人の個性によるものが強いのではないかと考えられる．

### 5.3.1 得られた被覆の大きさ

**表 2** は，実験の結果を示している．まず，1 つの歌集に生起するパターンの数は，数十万のオーダーであって，これを人手で吟味することは不可能であることが分かる．また，括弧内に示した数は，2 回以上生起したパターンに限定したときのパターンの数であるが，これでも，人手で検討できる数ではない．実験においては，パターンの候補集合である $\Delta$ として，この 2 回以上生起したパターンの集合を用いた．出力された被覆の大きさを，第 3 コラムに示した．たとえば，古今集においては，8,265 個のパターンの中から 191 個のパターンを被覆として抽出している．得られた被覆に含まれるパターンの数は，そのすべてを吟味するのに十分小さいといえる．

表 3 古今集のパターン
Table 3 FUSHI patterns from KOKINSHŪ.

| パターン | 説明 |
|---|---|
| *ければ*べらなり* | 古今集時代特有の助動詞べらなりの使用 |
| *ぞ*しかりける* | 係り結び |
| *こそ*りけれ* | 係り結び |
| *りせば*らまし* | 反実仮想 |
| *は*なりけり* | 気づきの表現 |

### 5.3.2 典型的なパターン

**表 3** は，古今集に対して得られたパターンのうち，アルゴリズムが出力した順，すなわち，基本的には，

$$u(\pi) - \frac{v(\pi)}{|L(\pi) \cap A|}$$

の値の大きい順に並べた際の最初の 5 つのパターンを示したものである．最初のパターン「*ければ*べらなり*」は，助動詞べらなりを含んでいるが，この助動詞は，主として古今集時代に用いられた助動詞として知られている．4 番目のパターン「*りせば*らまし*」は，反実仮想である．5 番目のパターン「*は*なりけり*」は，気づきの表現といえる．2 番目と 3 番目のパターンは，係り結びである．

このように，被覆として得られた 191 個のパターンのうち，出力の順位が上位のものは，典型的なパターンが多かった．他の歌集についても，おおむね同様の傾向が見られた．

### 5.3.3 自明でないパターン

被覆として得られたパターンを，出力の順位が上位のものから見ていくと，上位のものは，典型的ではあるものの，和歌の研究者にとっては自明なパターンであった．これに対し，中位のものにいくに従って，研究者にとって興味深いもの，すなわち，自明でないものが見られた．その例として，**表 4** に示すパターンがあげられる．歌集によって，かなり傾向の違いがあるようである．なお，ここにあげたパターンは，第 2 著者が選んだものであり，いうまでもなく，その判断には作業者の主観が入っている．

なお，得られたパターンの中には，ふしとは認め難いものも含まれていたが，1 章で最初に述べたように，得られたものすべてが有用であるようにすることは，本研究の目的ではない．実際，研究者が特定のパターンを除外したければ，それをパターンの候補集合 $\Delta$ から除くことも可能である．

### 5.3.4 歌集による差異

次に，歌集ごとに得られたパターンの生起頻度を，それ以外の歌集における生起頻度と比較した．**表 5** は，それぞれの歌集から得たパターンのうち出力順に並べ

た上位の5つのパターンについて，5つの歌集における生起頻度を示したものである．表から，以下のことが見てとれる．

(1) パターン「*ばかり*らむ*」は，古今集と新古今集には表れていない．

(2) パターン「*は*なりけり*」は，いずれの歌集にも表れている．特に，山家集にはこの表現が多い．

(3) 山家集の歌には，パターン「*まし*なりせば*」が多い．

表4 自明でないパターンの例
Table 4 Examples of non-obvious FUSHI patterns.

| 歌集 | パターン |
|---|---|
| 古今集 | *るらむ*しものを* |
| | *ならば*まし* |
| | *とも*らじ* |
| | *じ*とも* |
| | *は*にして* |
| 新古今集 | *こそは*め* |
| | *の*ければ* |
| | *ば*ばかりぞ* |
| 壬二集 | *とだに*らする*もがな* |
| | *も*ぬべし* |
| | *ても*かな* |
| | *きにけらしな*るまで* |
| 拾遺愚草 | *とや*けん* |
| 山家集 | *まし*かりせば* |
| | *いかに*らん* |
| | *ばやは*べき* |
| | *けりな*たれば* |
| | *さへ*かな* |
| | *と*ばかりぞ* |

(4) 拾遺愚草には，パターン「*こそ*りけれ*」がない．

(5) パターン「*や*るらむ*」は，古今集にはないが，他の4つの歌集には表れている．

これらは，時代や歌人の個性に関連した重要な特徴である可能性がある．たとえば，(2)，(3)は西行法師の個人的な好みの反映であるかもしれない．また，(5)はこのパターンが古今集時代には好まれなかったことを示しているかもしれない．このように，抽出されたパターンやその頻度の比較によって，研究者は，新たな問題意識を持つことができるのである．ここから先は文学の領域の問題であって，その解決には，文学の立場からの総合的な議論を待たなければならない．研究の主体であるべき和歌文学研究者の出番は，まさにここからなのである．

## 6. パターンに基づくテキストデータマイニングシステム

前章で述べた歌集からのパターン抽出の経験に基づき，著者らは，和歌文学研究支援のためのテキストデータマイニングシステムを作成した．**図2**にその概観を示す．このシステムは，パターン発見部とパターン照合部から成る．パターン発見部は，パターンの集合を出力するが，そのうちのあるものは，ユーザである研究者の思考を活性化し，仮説の生成を助ける．その仮説を検証するために，ユーザはパターン照合部を用いて，指定したパターンを含む和歌の用例を検索し，得られた用例を調べ，必要であれば仮説を修正する．

**表5** 5つの歌集から得られたパターンとその生起頻度．ここで，A，B，C，D，Eは，それぞれ，古今集，新古今集，壬二集，拾遺愚草，山家集を表す
Table 5 FUSHI patterns from five anthologies with frequencies, where A, B, C, D, and E denote KOKINSHŪ, SHINKOKINSHŪ, MINISHŪ, SHŪIGUSŌ, and SANKASHŪ, respectively.

| | パターン | A | B | C | D | E | | パターン | A | B | C | D | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | *ければ*べらなり* | 5 | 0 | 0 | 0 | 0 | D | *ばかり*らむ*(1) | 0 | 0 | 11 | 8 | 3 |
| | *ぞ*しかりける* | 8 | 1 | 0 | 0 | 3 | | *の*なりけり* | 19 | 30 | 39 | 19 | 49 |
| | *こそ*りけれ*(4) | 11 | 8 | 8 | 0 | 13 | | *らざりき*の* | 0 | 0 | 1 | 6 | 1 |
| | *りせば*らまし* | 5 | 2 | 0 | 0 | 4 | | *や*るらむ*(5) | 0 | 8 | 40 | 24 | 23 |
| | *は*なりけり*(2) | 20 | 26 | 26 | 11 | 52 | | *に*なるらむ* | 0 | 2 | 8 | 8 | 7 |
| B | *かりせば*まし* | 3 | 6 | 0 | 0 | 1 | E | *まし*なりせば* | 0 | 2 | 1 | 0 | 10 |
| | *の*にけるかな* | 4 | 11 | 2 | 1 | 4 | | *こそ*かりけれ* | 4 | 4 | 1 | 0 | 8 |
| | *は*なりけり*(2) | 20 | 26 | 26 | 11 | 52 | | *ならば*らまし* | 1 | 0 | 0 | 0 | 8 |
| | *こそ*りけれ*(4) | 11 | 8 | 8 | 0 | 13 | | *を*ふなりけり* | 1 | 0 | 0 | 0 | 7 |
| | *も*かりけり* | 4 | 11 | 8 | 5 | 7 | | *の*るなりけり* | 4 | 3 | 4 | 0 | 10 |
| C | *ばかり*るらむ* | 0 | 0 | 6 | 0 | 3 | | | | | | | |
| | *こそ*なりけれ* | 4 | 0 | 5 | 0 | 5 | | | | | | | |
| | *や*なるらむ* | 0 | 2 | 16 | 4 | 7 | | | | | | | |
| | *は*なりけり*(2) | 20 | 26 | 26 | 11 | 52 | | | | | | | |
| | *の*なりけり* | 19 | 30 | 39 | 19 | 49 | | | | | | | |

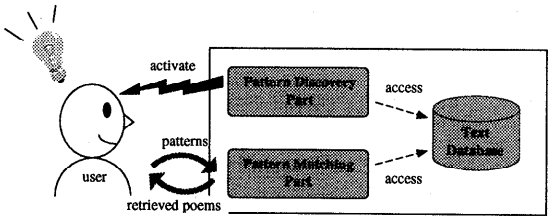



図 2 パターンに基づくテキストデータマイニングシステム
Fig. 2 Pattern-based text data mining system.

修正された仮説は，再び検証される．このようなプロセスを繰り返すことにより，研究者は新しい視点を開拓していく．

実用的には，パターン発見部によって出力されたパターンそのものでなく，少しだけ一般的（general）ないしは具体的（specific）なものが有用であることも多い．このため，ユーザは，パターンの近傍をブラウズできることが望ましい．そこで，このシステムは，パターンの集合上の半順序に関するハッセ図（Hasse diagram）をブラウズできるようなパターンブラウザの機能を持つ．著者らは，このシステムを用いて，新しいスタイルの和歌文学研究を進行中である．

## 7. おわりに

本論文では，和歌の集合からその特徴となるパターンを半自動的に抽出するための手法を示し，これを5つの歌集に適用した結果を述べた．抽出されたパターンの集合は，和歌の研究者にとって有用なものを多く含んでいた．また，このような研究を支援するためのテキストデータマイニングシステムを示した．現在，そのプロトタイプシステムを用いて研究を続けている．

もちろん，本来このようなパターンは，研究者自身が歌集を読みこなすことによって着想すべきものである．しかし，計算機により自動抽出されたパターンによって，研究者の発想が活性化され，新たな視点を見出す端緒となることが期待できる．すなわち，本論文で提案した手法は，和歌文学研究者を主体とした研究を支える有用なツールとなりうるのである．

本研究で扱った，歌集からのふしの抽出の問題に関しては，今後の課題として次のようなことが考えられる．

(1) 付属語列の収集．
(2) 付属語列を句末に限定したことによる失敗の改善．
(3) パターンの「重要」性に関する別の尺度の導入．

(1) に関しては，実験では，付属語列の集合として，助詞・助動詞および活用語尾から成るものに限定したデータを用いたが，もちろん，これでは十分ではない．たとえば，2.4 節で示した「のまにまに」などを付属語列として扱えるようにすることが必要である．そこで，このような付属語列を網羅的に収集する方法が課題となる．(2) は，句の途中の付属語列がふしに関与している場合の問題である．ふしに関与する付属語列のうち句の途中に頻出するものとしては，係り結びをつくる「ぞ」「こそ」などが多いようであり，これについては改善の余地があろう．(3) に関しては，本研究では，パターンの「重要」性の尺度として，MDL 原理に基づくものを採用したが，もちろん，ほかの尺度を用いることも考えられる．この問題は，どのようなパターンを得たいかということに強く依存している．たとえば，表 5 で示したパターンの中では，ある歌集に多く生起するが，別の歌集では生起しないようなものが有用であった．そこで，このようなものだけを取り出すような尺度を与えることも考えられよう．また，5.3.3 項で示した自明でないパターンをうまく得るために，非自明性（nonobviousness）に関する尺度を与えることも，非常に興味深い課題である．

和歌文学研究において，和歌の類似性に着目することは重要である [19)~21)]．和歌の類似性としては，本論文で論じた「歌語に関する類似性」と「ふしに関する類似性」のほかにも，様々なものが考えられよう．その1つとして，現在著者らは，「和歌を単なる文字の連鎖と見なした場合に共通した文字列を多く含む」という類似性を考えている [22)]．このような意味で類似した和歌は，同一の歌であったものが，伝来の過程で本文が改変されたものであるかもしれない．そこで，伝来の過程で異同を生じた理由を追求することにより，和歌史的考察にまで発展することが期待できる．あるいは，このような類似歌は，本歌取りといって古歌を踏まえて新たに歌を詠んだものであるかもしれない．この場合には，本歌取りにおいてどのように古歌を踏まえているかを分析することを通じて，表現技法に関する知見を得ることができよう．いずれにせよ，類似歌の発見を契機にして，新たな問題意識が生まれることになる．著者らは，古今集と新古今集との間の類似歌抽出の実験を行い，最新かつ代表的な注釈書には記載のない本歌取りが指摘できることを示している [22)]．

大量のデータからそこに潜む規則やパターン，類似性などを半自動的に抽出する技法は，国文学の研究においても有用である．現状では，国文学分野のデータベース化は，欧米に比べ立ち後れている [23)] が，今後その作業が進めば，このような研究の重要性は，ますます顕在化するであろう．

**謝辞** 九州大学の有川節夫教授，今西裕一郎教授，総合研究大学院大学の及川昭文教授，村上征勝教授には，本研究に関し貴重なご助言をいただいた．また，九州大学の有村博紀助教授，篠原歩助教授には，パターン抽出手法に関して熱心にご議論いただいた．佐竹昭廣教授をはじめとする国文学研究資料館の方々には，本研究の初期段階において，国文学研究における計算機利用の在り方に関して，非常に示唆的なお話をおきかせいただいた．ここに深謝いたします．

## 参考文献

1) 八村広三郎：人文科学とデータベース，情報処理学会誌，Vol.38, No.5, pp.377–382 (1997).
2) 中村康夫：国文学研究資料館のデータベース—特に国文学論文目録データベースについて，人文学と情報処理，No.2, pp.61–67 (1993).
3) 安永尚志：日本古典文学本文データベース形成とデータ記述文法，情報処理学会「人文科学とコンピュータ」研究報告，No.91-CH-8 (1991).
4) 安永尚志：日本古典文学の本文データベース，情報処理学会誌，Vol.35, No.7, pp.642–650 (1994).
5) 村上征勝，上田英代，樺島忠夫，今西裕一郎，上田裕一：単語情報に基づく源氏物語の計量分析，情報処理学会研究報告，No.95-CH-26, pp.55–60 (1995).
6) Brāzma, A., Ukkonen, E. and Vilo, J.: Discovering unbounded unions of regular pattern languages from positive examples, *Proc. 7th International Symposium on Algorithms and Computation (ISAAC '96)*, pp.95–104 (1996).
7) Rissanen, J.: Modeling by the shortest data description, *Automatica*, No.14, pp.465–471 (1978).
8) Fayyad, U.: From data mining to knowledge discovery: an overview, *Advances in knowledge discovery and data mining*, Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds.), pp.1–34, AAAI Press (1972).
9) Yamasaki, M., Takeda, M., Fukuda, T. and Nanri, I.: Discovering characteristic patterns from collections of classical Japanese poems, *Proc. 1st International Conference on Discovery Science (DS '98)*, pp.129–140 (1998).
10) 福田智子：藤原元真の作歌法—物名歌人の横顔，香椎潟，No.44 (1999). to appear.
11) 山本 靖：和歌集テキストにおける未知語品詞推定および形態素辞書構築支援への利用，修士論文，奈良先端科学技術大学院大学情報科学科 (1996).
12) Angluin, D.: Finding patterns common to a set of strings, *Proc. 11th Annual Symposium on Theory of Computing*, pp.130–141 (1979).
13) Gold, E.M.: Language identification in the limit, *Information and Control*, No.10, pp.447–474 (1967).
14) Valiant, L.: A theory of the learnable, *Comm. ACM*, Vol.11, No.27, pp.1143–1142 (1967).
15) Shinohara, T.: Polynomial time inference of pattern languages and its applications, *Proc. 7th IBM Symposium on Mathematical Foundations of Computer Science*, pp.191–209 (1982).
16) Arimura, H., Shinohara, T. and Otsuki, S.: Finding minimal generalizations for unions of pattern languages and its application to inductive inference from positive data, *Proc. 11th Annual Symposium on Theoretical Aspects of Computer Science (STACS '94)*, pp.649–660 (1994).
17) Brāzma, A., Jonassen, I., Ukkonen, E. and Vilo, J.: Discovering patterns and subfamilies in biosequences, *Proc. 4th International Conference on Intelligent Systems for Molecular Biology*, pp.24–43 (1996).
18) Garey, M. and Johnson, D.: *Computers and intractability: A guide to the theory of NP-Completeness*, W.H. Freeman and Company, San Francisco (1979).
19) 福田智子：藤原兼家六十賀和歌をめぐって—正保版歌仙家集本『兼盛集』と西本願寺本『能宣集』，国語国文，Vol.62, No.12, pp.1–14 (1993).
20) 福田智子：「大原の山」から「大原の里」へ，語文研究，No.77, pp.1–10 (1996).
21) 福田智子：平祐挙の歌—一条朝和歌の一側面，和歌文学研究，No.75, pp.11–21 (1997).
22) 山崎真由美，竹田正幸，福田智子，南里一郎：和歌データベースからの類似歌の自動抽出，情報処理学会「人文科学とコンピュータ」研究報告，Vol.98, No.97, pp.57–64 (1998).
23) 長瀬真理：文学データベース—急がれる総合的な環境整備，情報処理学会誌，Vol.38, No.5, pp.397–400 (1997).

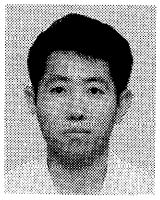
(平成10年 8 月 31 日受付)
(平成10年 11 月 9 日採録)

**竹田 正幸**（正会員）

1964年生．1987年九州大学理学部数学科卒業．1989年同大学大学院総合理工学研究科情報システム学専攻修士課程修了，同年より同大学工学部電気工学科助手．1996年より同大学大学院システム情報科学研究科情報理学専攻助教授，現在に至る．博士（工学）．パターン照合アルゴリズム，テキスト圧縮，機械学習，機械発見，情報検索，自然言語処理などの研究に従事．人工知能学会，日本ソフトウェア科学会，和歌文学会各会員．

**福田 智子**

1964年生．1987年福岡女子大学文学部国文学科卒業．1992年九州大学大学院文学研究科国語学・国文学専攻修士課程修了．1997年同専攻博士後期課程単位取得退学．1994年より福岡女学院大学非常勤講師，現在に至る．平安朝文学，特に和歌を中心に研究する．和歌文学会，中古文学会，西日本国語国文学会各会員．



**南里 一郎**

1966年生．1990年九州大学文学部文学科卒業．1995年同大学大学院文学研究科国語学・国文学専攻修士課程修了．1997年同専攻博士後期課程中途退学．同年より純真女子短期大学国文科講師，現在に至る．平安時代語・鎌倉時代語に興味を持つ．国語学会，西日本国語国文学会各会員．

**山崎真由美**（学生会員）

1974年生．1997年九州工業大学工学部電気工学科卒業．同年，九州大学大学院システム情報科学研究科情報理学専攻修士課程入学，現在に至る．機械学習，機械発見などに興味を持つ．